ISSP 2024 - 13th International Seminar on Speech Production
13-17 May 2024, Autrans, France

Some Effects of Frame Rate on Gesture Detection in Tongue Ultrasound

Pertti Palo, Steven M. Lulich

Indiana University Bloomington
pertti.palo@taurlin.org, slulich@indiana.edu

Abstract

We study how decreasing ultrasound frame rate affects automated speech gesture detection. The gesture detection is performed on Pixel Difference contours with simulated stepping down of the frame rate from ≈122Hz down to ≈17Hz. We report how this affects the number of peaks detected and the accuracy of peak locations for Pixel Difference calculated using six different vector norms. The results point to a steady degradation of detection results as the frame rate is decreased.

Keywords: speech timing, speech gestures, sampling frequency, automated methods

1. Introduction

While tongue ultrasound is widely used in speech research and related areas, the analysis is often limited to selecting points of interest based on acoustic segmentation and then analysing the corresponding frames by extracting tongue splines. With the advent of reliable automated splining methods (Wrench and Balch-Tomes 2022), and in the case of using holistic image based methods, we are no longer limited to basing the analysis on comparing single sample points. Instead, we can analyse articulation as a (almost) continuous function of time (Palo 2019; Al-Tamimi and Palo 2023). In time domain analysis, the sampling frequency or frame rate of the data becomes an important factor that can limit the analysis we are able to perform (Palo and Lulich 2023).

Palo and Lulich (2023) used a method called Pixel Difference (PD) for speech gesture analysis. PD evaluates the overall change in an ultrasound image sequence by interpreting the images as vectors and calculating the distance between consecutive images as a vector norm (Palo 2019). Similar methods have been used by, for example, Drake, Schaeffler, and Corley (2013) and Raeesy, Baghai-Ravary, and Coleman (2011). **Figure 1** demonstrates PD and the effect of lower frame rates on this type of analysis.

To state this problem broadly, we are interested in what the limit frequency is for speech articulation gestures to be detectable in articulatory data. More specifically, we will concentrate on tongue ultrasounds. The simple answer to this type of question in signal processing comes from the Nyquist-Shannon sampling theorem and states that to detect a signal without aliasing artefacts we need a sampling frequency that is at least double the frequency of the signal (Shannon 1949). However, we are going to need to do better than just detect a signal at the frequency of interest.

In a related study, Wrench and Scobbie (2008) used data from two different ultrasound systems. They sampled extracted tongue contours along two directions from the ultrasound probe origin to produce graphs of contour movement in the root and tip regions of the tongue. They show that 60 Hz data (produced by de-interlacing 30 Hz video ultrasound data) produces analysis results that are on par with those produced by a 100 Hz system showing a clear-to-the-eye difference between the /ele/-gesture in "Pay Laver" vs. "Pale Eva" in both types of data. On the other hand, spectral analysis of X-ray microbeam point tracking data shows that most of its information is in the under 12 Hz band as shown by correlation analysis with acoustic data (Goldstein 2019).

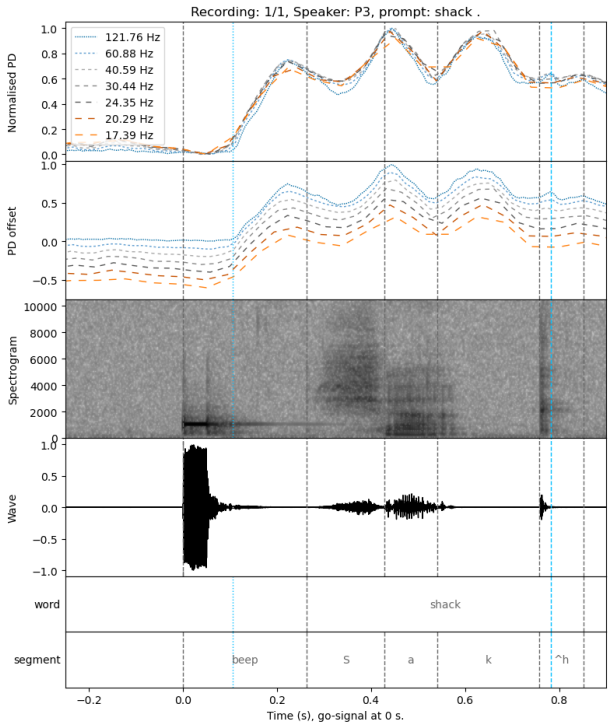

Figure 1: *Effects of downsampling on PD. Top panel shows PD curves resulting from no downsampling (121.76 Hz) up to a factor of 7 (17.39 Hz). The second panel shows the same data offset for clarity of individual curves. Bottom panels are the spectrogram, waveform (the go-signal beep is the part with the largest amplitude) and word with phonological segmentation. The vertical dotted blue line marks movement onset on the original data, and the vertical dashed blue line marks a gesture peak associated with aspiration of /k/.*

Analysing gesture timing without anatomical reference points is quite different from analysing contour movement and despite some commonalities - analysing flesh-point tracking like X-ray microbeam is different from PD analysis of image sequence data like ultrasound. For detailed analysis of speech timing from time series, we are going to need to be able to iden-

10.21437/issp.2024-59

tify minima and maxima at a good enough accuracy and to not lose any that are produced by the gestures. In **Figure 1** we can see why the latter are a concern. Looking at articulatory onset (marked with vertical dotted blue line), we can see that at lower sampling frequencies we would not be able to identify it as accurately, which would quickly produce problems in study designs that need high statistical power. Furthermore, the gesture peak marked with a vertical dashed blue line disappears completely as sampling frequency goes down.

Continuing our recent work (Palo and Lulich 2023), we seek to empirically determine what the sampling frequency of tongue ultrasound needs to be in order for automatic peak detection to be able to find a believable number of gesture peaks in an utterance. To do so, we explore the effect of two variables on peak detection: the sampling frequency and the vector norm (a type of l_n^p-norm) used to calculate PD.

2. Materials

The data is a sample of 174 single-word utterances of a delayed naming experiment. The words were single-syllable lexical English words with a word final plosive ([p, t, k]) and an onset consonant ranging from none to /CCC/. The data was recorded at 121.76 fps in the mid-sagittal plane synchronised with audio. For details, please see Experiment 2, Participant 3 in Palo (2019). This speaker's data has good tongue surface visibility and provides a good baseline for this proof-of-concept study.

3. Methods

3.1. Downsampling

The original data is downsampled by a factor ranging from 2 to 7. For a downsampling factor of n this is done by using only every n^th frame in the ultrasound data for analysis. For example, for a factor of 3, we use frames [1, 4, 7, 10, ...] as the analysed data. Since the original data was recorded at 121.76 fps, this gives the sampling frequencies shown in **Figure 1**: 60.88, 40.59, 30.44, 24.35, 20.29, and 17.39 Hz.

3.2. Vector norms

Vector lp norms – or more precisely l_n^p-norms – can be defined as shown in Equation 1. In our case p is the order of the norm, n length of the vector or size of the ultrasound frame in pixels, and x_i are the individual elements of the vector, which in PD are evaluated as differences between corresponding pixels in consecutive frames.

$$l_n^p = \begin{cases} \sum_{i=1}^{n} \frac{|x_i|}{1+|x_i|}, & p = 0 \\ \sum_{i=1}^{n} |x_i|^p, & 0 < p < 1 \\ \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}, & 1 \leq p < \infty \\ \max(|x_i|), & p = \infty \end{cases} \quad (1)$$

Since the parameter n is defined by the number of pixels in the analysed frames, we will use the simpler notation of lp in the rest of this paper. We chose to use the norms $l0.5$ $l1$, $l2$, and $l5$ to provide a sample around $l1$ and $l2$, which we have used previously, and $l0$ and $l\infty$ because they are the limits of the range of p.

3.3. Peak detection

Gestures were identified automatically with the function `scipy.signal.find_peaks` from the SciPy software package (Virtanen et al. 2020). We used three parameters – `distance`, `width`, and `prominence` – to tune the peak selection process and produce reasonable accuracy in identifying actual gesture peaks. The process was guided by observing the results on a test set of 10 recordings for norms $l0.5$ $l1$, $l2$, and $l5$. The recordings were the first 10 in the data set.

A conservative lower limit for the gesture interval (parameter `distance`) was estimated from the data of Jacewicz, Fox, and Wei (2010). They report a high limit of approximately 6.7 syllables/second for speech rate (see Figure 1 in Jacewicz, Fox, and Wei (2010)). Given that syllables can be expected to have at least two gestures associated with them, we arrive at a lower bound of $t_{lower} = \frac{1}{2*6.7} \approx 0.075\ s$ for the interval between gestures. This interval length was adapted for downsampling by scaling it accordingly and rounding up.

The `width` parameter was chosen as 1 (meaning a peak with a width of 1 sample halfway down its prominence value was accepted as valid). The test set would have merited using a higher value if we were only interested in getting the best results for that set. However, using a higher value would make peak detection deteriorate very fast with downsampling as time spanned by 3 frames expands. We are still going to see degradation of the results when the sampling frequency gets close to the Nyquist frequency. This is actually desirable because the articulatory gestures are not sinusoidal signals, and in order to analyse them we need better time resolution than that required by the Nyquist frequency condition.

Finally, `prominence` was selected by stepping its value within the set (0.005, 0.01, 0.02, 0.03, 0.04). The last value was found to exclude peaks in the test set that we did not want to exclude, and so we used the value 0.03 for the `prominence`. It should be noted that while the individual parameters behave occasionally in an unintuitive manner, on the whole the way that `scipy.signal.find_peaks` works provides a very intuitive and easy to use way of identifying the peaks we are interested in.

3.4. Choosing the Period of Interest

Lower limit of the Period of Interest (POI) was set at 58 ms from the beginning of the go-signal (50 ms long 1 kHz sinusoidal beep) after Palo (2019) based on the minimal reaction times calculated by Chiu and Gick (2014). The upper limit of the POI was set the length of a gesture interval after end of the word to include the possible plosive release gesture in the analysis.

We used the chosen inter-gesture interval to extend the POI from the end of the utterance-final burst segment's *beginning* to account for cases where the plosive produced only a short release burst, and thus the acoustic boundary is at times already before the release gesture peak or very close to it.

4. Results

Our results are illustrated in **Figures 2-4**. Downsampling causes the number of detected peaks (**Figure 2**) to mainly decline for all of the norms with $l1$, $l2$, and $l5$ showing the best stability. However, the sample-by-sample peak number ratio distributions show that in some cases downsampling first increases the number of detected peaks as evident in that some distribution tails in **Figure 3** are above 1. This effect is strongest in $l5$ and $l\infty$.

Figure 4 shows that the peak position errors increase for

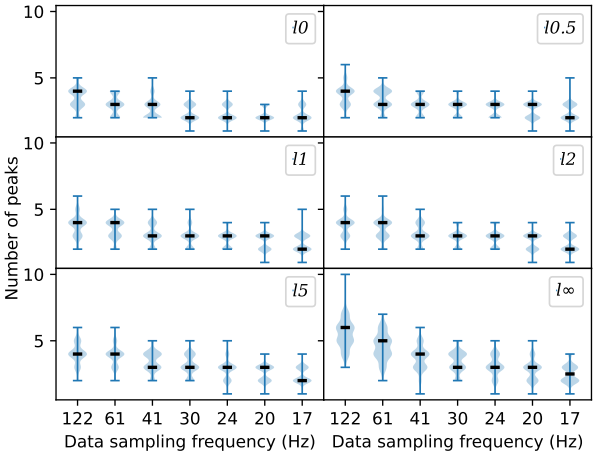

Figure 2: *Distributions of number of peaks detected in each sample. Black bars mark distribution medians.*

all of the norms while $l1$ and $l2$ behave the best in this respect. In this figure we relate the position errors to the limit set above in Section 3.3 for the minimum time between gestures: $t = 0.075$ s. All of the error distribution tails cross the limit already at 41 fps. At 30 fps and below there are more than outliers above the limit for each norm. None of the distribution medians cross the limit before 17 fps.

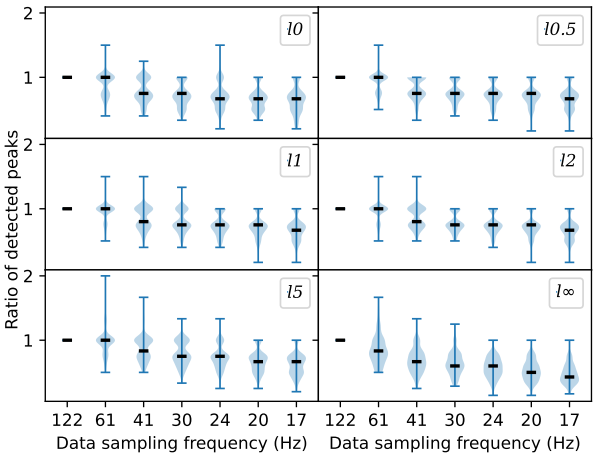

Figure 3: *Distributions of ratio of peaks detected in each sample, compared to those in the original data (122 Hz). Black bars mark distribution medians.*

5. Discussion

A necessary caveat on our results is that the approach we have taken is not exactly the same as using ultrasound with a lower frame rate. This is because the analysis here achieves a lower frame rate by dropping frames. As such it remains unclear if a longer frame acquisition time will affect the quality as well. This seems likely as longer frame acquisition means that the likelihood of within-frame movement artefacts increases.

As for the speech materials analysed, the data comes from only a single speaker. Further, it should be noted that our selection of phonetic content is limited. We did not have any flaps, taps, or trills in the test dataset. Of these flaps and taps are likely to be the fastest tongue gestures and should be included in a more comprehensive analysis. Trills on the other hand are held longer in terms of the whole tongue, and while they have a dynamic target, attaining the target can be imaged with a lower frequency than two times the trill frequency.

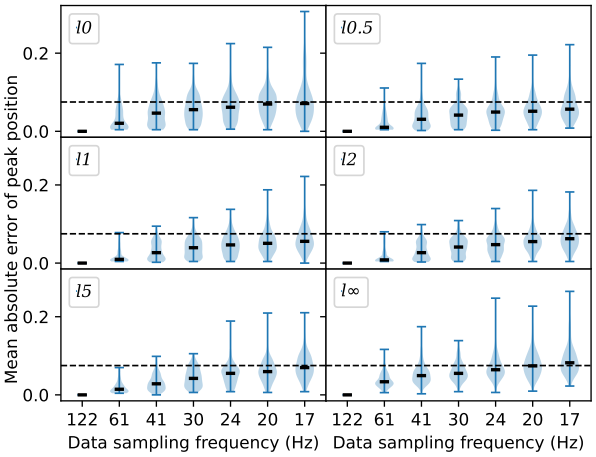

Figure 4: *Distributions of time accuracy of peak detection compared to the original data. Black bars mark distribution medians and the dashed line marks the lower limit of the time between gestures t=0.075 s used in the peak detection.*

6. Conclusion

The results show that quality of automatic peak detection degrades steadily with dropping of the sampling frequency. There does not seem to be any kind of division into two regions where the results would be good down to a given sampling frequency and then sharply change after that. Instead, the results point to a conclusion that a higher sampling frequency is always desirable.

There is no clear winner in terms of the used norms either. There is clear indication, however, that the limit norms – $l0$ and $l\infty$ should not be used. Rather, if there is an optimal norm or norm region, it will probably be somewhere close to $l1$ and $l2$.

As for the method itself, the results do show that useful, actionable information can be gained by this type of analysis. In particular, this study provides a reason to prefer high frame rates when using automated gesture detection. Before analysing a larger and more varied data set, the conclusion about lower frame rates of data must remain only a tentative caution.

7. Acknowledgements

Pertti Palo's work has been funded by a post-doctoral grant from the Emil Aaltonen foundation via the Post-Doc Pool of Finland.

8. References

Chiu, C and B. Gick (2014). "Startling Speech: Eliciting Prepared Speech Using Startling Auditory Stimulus". In: *Frontiers in Psychology* 5.1082.

Drake, E., S. Schaeffler, and M. Corley (2013). "ARTICULATORY EVIDENCE FOR THE INVOLVEMENT OF THE SPEECH PRODUCTION SYSTEM IN THE GENERATION OF PREDICTIONS DURING COMPREHENSION". In: *Architectures and Mechanisms for Language Processing (AMLaP)*. Marseille.

Goldstein, Louis (2019). "The Role of Temporal Modulation in Sensorimotor Interaction". In: *Frontiers in Psychology* 10. DOI: 10.3389/fpsyg.2019.02608. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.02608/full (visited on 04/16/2024).

Jacewicz, Ewa, Robert Allen Fox, and Lai Wei (2010). "Between-Speaker and within-Speaker Variation in Speech Tempo of American English". In: *The Journal of the Acoustical Society of America* 128.2, pp. 839–850. DOI: 10.1121/1.3459842.

Palo, P. (2019). "Measuring Pre-Speech Articulation". PhD thesis. Edinburgh: Queen Margaret University.

Palo, P. and S. M. Lulich (2023). "Improving Signal-to-Noise Ratio in Ultrasound Video Pixel Difference". In: *The Journal of the Acoustical Society of America* 153.3_supplement, A373. DOI: 10.1121/10.0019222.

Raeesy, Z., L. Baghai-Ravary, and J. Coleman (2011). "Parametrising Degree of Articulator Movement from Dynamic MRI Data". In: *12th Interspeech*, pp. 2853–2856.

Shannon, C.E. (1949). "Communication in the Presence of Noise". In: *Proceedings of the Institute of Radio Engineers* 37.1, pp. 10–21. DOI: 10.1109/JRPROC.1949.232969. URL: https://ieeexplore.ieee.org/document/1697831 (visited on 04/09/2024).

Al-Tamimi, J. and P. Palo (2023). "Dynamics of the Tongue Contour in the Production of Guttural Consonants in Levantine Arabic". In: *International Conference of Phonetic Sciences (ICPhS 2023)*. Prague.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wrench, A. and J. Balch-Tomes (2022). "Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut". In: *Sensors* 22, p. 1133. DOI: doi:10.3390/s22031133.

Wrench, A. and J. M. Scobbie (2008). "High-Speed Cineloop Ultrasound vs. Video Ultrasound Tongue Imaging: Comparison of Front and Back Lingual Gesture Location and Relative Timing." In: *Proceedings of ISSP 2008 - 8th International Seminar on Speech Production*.